BBA - Molecular Basis of Disease xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - Molecular Basis of Disease

journal homepage: www.elsevier.com/locate/bbadis

Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis☆

ShaoPeng Wang, YuDong Cai*

School of Life Sciences, Shanghai University, Shanghai 200444, People's Republic of China

ARTICLE INFO

Keywords:
Cancer prediction
Support vector machine
Gene ontology
KEGG
Monte-Carlo feature selection

ABSTRACT

Cancers are regarded as malignant proliferations of tumor cells present in many tissues and organs, which can severely curtail the quality of human life. The potential of using plasma DNA for cancer detection has been widely recognized, leading to the need of mapping the tissue-of-origin through the identification of somatic mutations. With cutting-edge technologies, such as next-generation sequencing, numerous somatic mutations have been identified, and the mutation signatures have been uncovered across different cancer types. However, somatic mutations are not independent events in carcinogenesis but exert functional effects. In this study, we applied a pan-cancer analysis to five types of cancers: (I) breast cancer (BRCA), (II) colorectal adenocarcinoma (COADREAD), (III) head and neck squamous cell carcinoma (HNSC), (IV) kidney renal clear cell carcinoma (KIRC), and (V) ovarian cancer (OV). Based on the mutated genes of patients suffering from one of the aforementioned cancer types, patients they were encoded into a large number of numerical values based upon the enrichment theory of gene ontology (GO) terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. We analyzed these features with the Monte-Carlo Feature Selection (MCFS) method, followed by the incremental feature selection (IFS) method to identify functional alteration features that could be used to build the support vector machine (SVM)-based classifier for distinguishing the five types of cancers. Our results showed that the optimal classifier with the selected 344 features had the highest Matthews correlation coefficient value of 0.523. Sixteen decision rules produced by the MCFS method can yield an overall accuracy of 0.498 for the classification of the five cancer types. Further analysis indicated that some of these features and rules were supported by previous experiments. This study not only presents a new approach to mapping the tissue-of-origin for cancer detection but also unveils the specific functional alterations of each cancer type, providing insight into cancer-specific functional aberrations as potential therapeutic targets. This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

1. Introduction

Cancer is regarded as a malignant proliferative disease that can occur in many tissues and organs in humans [1,2]. As a systemic disease, the symptoms of cancer are not restricted to the sites of tumorigenesis [2]. The proliferative, invasive and metastatic characteristics of cancer have been associated with a high mortality rates [3–5]. In 2012, 14.1 million new cancer cases were diagnosed, and at the same time, approximately 8.2 million people died of such disease. Based on statistical prediction, by 2025, more than 19.3 million people may be diagnosed with cancer, demonstrating that cancer is one of the major threats to human life [6].

It is well known that the early diagnosis of cancers can greatly increase the chances of successful treatment and survival of patients. Cell-free DNA (cfDNA) has been recognized as a potential non-invasive cancer biomarker since the discovery of *TP53* mutations in the urinary sediments of bladder cancer patients and the detection of mutated *RAS* gene in the blood of cancer patients [7–9]. The liquid biopsy of cfDNA in plasma or serum could avoid the need for tumor tissue biopsies and allow the cfDNA to be monitored during the progression and the treatment of cancers. Information about the tissue-of-origin from the liquid biopsies are important for locating and diagnosing the primary cancers early but require knowledge of the cancer-specific or tissue-specific variations. For example, tissue-specific DNA methylation, cell-specific nucleosome occupancy pattern and cancer-specific mutation signatures are now available to characterize these biopsies [10–14].

☆ This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

* Corresponding author.

E-mail address: caiyudong@staff.shu.edu.cn (Y. Cai).

https://doi.org/10.1016/j.bbadis.2017.12.026
Received 17 October 2017; Received in revised form 4 December 2017; Accepted 15 December 2017

Meanwhile, specific mutation patterns have been identified as genetic characteristics to identify tumor types. For example, *ALK* gene rearrangement and its over-expression have been confirmed to be associated with non-small cell lung cancer and anaplastic large cell lymphoma [15]. Therefore, *ALK* gene and its expression products may serve as a core biomarker for the diagnostic and prognostic evaluation of these two cancer types [16]. The identification and clinical application of confirmed tumor genetic markers (mutation patterns) provide a new method to diagnose tumor types and distinguish them from each other. However, these mutated genes or gene products do not function in isolation but interact with each other in cellular networks and processes [17]. Thus, it is a more robust approach to identify the core unique characteristics of various tumor types at the level of biological processes rather than mutation signatures.

Unlike genes that are represented by specific gene names and symbols in computational biology, the biological processes are described by multiple bioinformatics initiatives based upon different point cuts. There are two core bioinformatics initiatives that contribute to the identification and description of functional biological processes and pathways in humans and across different species: gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [18,19]. GO compiles bioinformatics initiative describing genes and gene products by clustering their interactions with each other and annotating their respective contribution to certain biological processes [18,20]. In addition, the KEGG pathways provide a new approach to investigate biological processes. KEGG pathway terms cluster the functional genes into identified functional pathways, reflecting the real contribution of such genes to the living organism [19]. Therefore, during the identification of core unique biological factors in different tumor types, both GO terms and KEGG terms evaluate the differences from the point of view of integrated biological processes in a more comprehensive and convincing manner.

In this study, we applied a pan-cancer analysis to five different types of cancers: (I) breast cancer (BRCA), (II) colorectal adenocarcinoma (COADREAD), (III) head and neck squamous cell carcinoma (HNSC), (IV) kidney renal clear cell carcinoma (KIRC), and (V) ovarian cancer (OV). We obtained the somatic mutations found in these five cancer types from TCGA (The Cancer Genome Atlas) through the cBio cancer genomics portal [21–23]. Based upon the obtained mutated genes, patients with each aforementioned cancer type were encoded into a large number of numerical values using the enrichment theory of GO terms and the KEGG pathway [24–27]. The Monte-Carlo Feature Selection (MCFS) method [28] was adopted to analyze the GO term features and KEGG pathway features, yielding a feature list and sixteen decision rules. This feature list was used for the incremental feature selection (IFS) method to discover the most appropriate features for building the optimal classifier using the classic machine learning algorithm, support vector machine (SVM) [29,30], which could distinguish the five types of cancers with the best performance. This optimal SVM-based classifier provided a Matthews correlation coefficient value of 0.523 and an overall accuracy of 0.619. With regard to the sixteen decision rules, they can provide more clues to understanding the specific functional alterations of each cancer type than the classifier mentioned above, although it yielded a low overall accuracy of 0.498. Finally, important GO terms and KEGG pathways involved in the decision rules and optimal SVM-based classifier were extensively analyzed according to previous experimental results. Our study not only shed light on the mapping of the tissue-of-origin for cancer detection but also classified the functional alteration signatures of the five types of cancers, providing insight into the cancer-specific functional aberrations as potential therapeutic targets.

Table 1
The number of samples in each of the five cancer types.

Cancer type	Full name	Number of samples
BRCA	Breast cancer	513
COADREAD	Colorectal adenocarcinoma	499
HNSC	Head and neck squamous cell carcinoma	306
KIRC	Kidney renal clear cell carcinoma	473
OV	Ovarian cancer	456
Total	–	2247

2. Materials and methods

2.1. Materials

The mutational data in different types of cancers were downloaded from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/) [23], which contained the mutations in eleven cancer types. Because many cancer types only have very few samples compared with others, cancer types with less than 300 samples were excluded. The remaining five major cancer types included (I) BRCA, (II) COADREAD, (III) HNSC, (IV) KIRC, and (V) OV. The numbers of samples for these five cancer types are listed in Table 1.

2.2. The functional profiles of mutations

There have been many ways to describe a protein, such as the protein sequence based features [31] and secondary structure based features [32]. But the most direct one was the functional annotation of a protein from databases like GO and KEGG. There were limitations of direct binary annotation of whether a protein had a specific function. Such binary functional features will be very sensitive to the mis-annotations in the database. Therefore, the enrichment scores which considered the significance of overlap between a gene set and a GO or KEGG function in the genome background, will be more robust and give a quantitative measurement of function rather than a binary qualitative measurement [33]. In this study, we used the GO and KEGG enrichment scores [24–27] of mutated genes to measure the similarity of the functional effects caused by mutations between cancer patients.

2.2.1. GO enrichment score

For a given cancer patient p and one GO term GO_j, let G_{GO} denote the set of annotated genes of GO_j and $G(p)$ denote the set of mutated genes of cancer patient p. The GO enrichment score of p and GO_j is defined as the hypergeometric test P value [24–27,34–37] on $G(p)$ and G_{GO}, which can be computed with the following equation:

$$S_{\mathrm{GO}}(p, GO_j) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right) \tag{1}$$

where N and M denote the total number of human genes and the number of genes in G_{GO}, respectively; n and m represent the number of mutated genes in $G(p)$ and the number of genes both in $G(p)$ and G_{GO}, respectively. The higher the score, the stronger the functional effects of mutations in patient p on the GO term GO_j are. Overall, 19,997 GO terms were used in this study, inducing 19,997 GO enrichment scores for each cancer patient.

2.2.2. KEGG enrichment score

A similar approach was adopted to define the KEGG enrichment score, which can measure the associations between patients and KEGG pathways. Let G_{KEGG} denote the set of annotated genes of one KEGG pathway K_j. The KEGG enrichment score of p and K_j is defined as the hypergeometric test P value [24–27,34–37] on $G(p)$ and G_{KEGG}. This score can be calculated using the following equation: